Free Energy Wells and Overlap Gap Property in Sparse PCA

Ilias Zadik

CDS Moore-Sloan Postdoctoral Fellow,
Center for Data Science (CDS), NYU

Joint work with
Gérard Ben Arous (NYU) , Alexander S. Wein (NYU)

Simons workshop 2020: Computational Phase Transitions

September 25, 2020

The Sparse PCA Model

Setup: *rank-1 plus noise*

Let $x \in \{0,1\}^n$ be a **binary** k-**sparse** vector, $\lambda > 0$ the SNR, $k = o(n)$.

- For "additive noise" $W \in \mathbb{R}^{n \times n}$ *GOE matrix*
  i.e. W_{ii} i.i.d. $\mathcal{N}(0, 2/n)$, $W_{ij} = W_{ji}$ i.i.d. $\mathcal{N}(0, 1/n)$,
- we observe

$$Y = \lambda x x^\top + W.$$

The Sparse PCA Model

Setup: *rank-1 plus noise*

Let $x \in \{0,1\}^n$ be a **binary** k-**sparse** vector, $\lambda > 0$ the SNR, $k = o(n)$.

- For "additive noise" $W \in \mathbb{R}^{n \times n}$ *GOE matrix*
  i.e. W_{ii} i.i.d. $\mathcal{N}(0, 2/n)$, $W_{ij} = W_{ji}$ i.i.d. $\mathcal{N}(0, 1/n)$,
- we observe

$$Y = \lambda x x^\top + W.$$

Minimum λ so that, given Y, x can be recovered (efficiently) with probability tending to 1 as $n \to +\infty$ **(w.h.p.)**.

The Sparse PCA Model - Literature

Vast Literature. [Baik et al'05], [Johnstone, Lu '09], [Montanari et al '15], [Banks et al'18], [Ding et al'19], [Barbier et al'19], [Gamarnik et al '19]

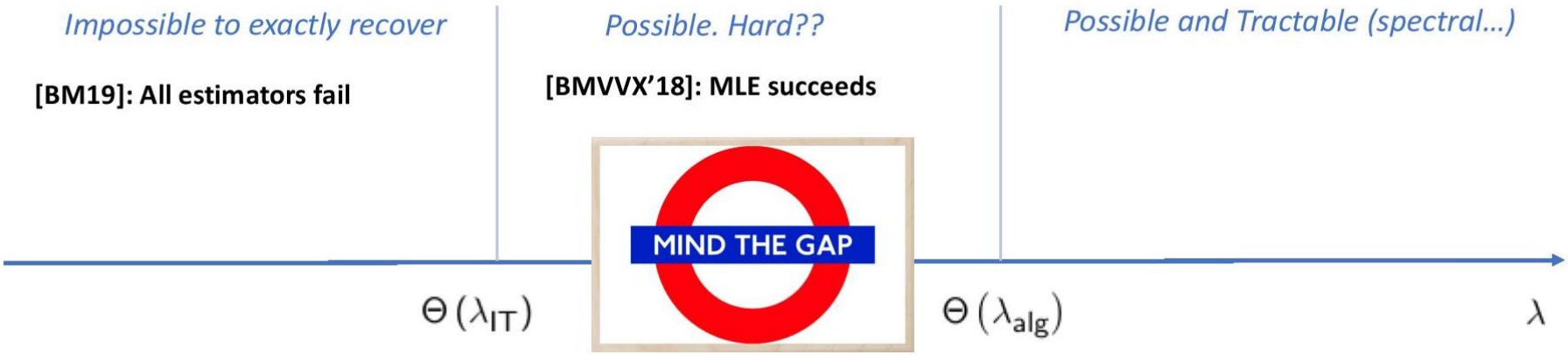


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$

The Sparse PCA Model - Literature

Vast Literature. [Baik et al'05], [Johnstone, Lu '09], [Montanari et al '15], [Banks et al'18], [Ding et al'19], [Barbier et al'19], [Gamarnik et al '19]

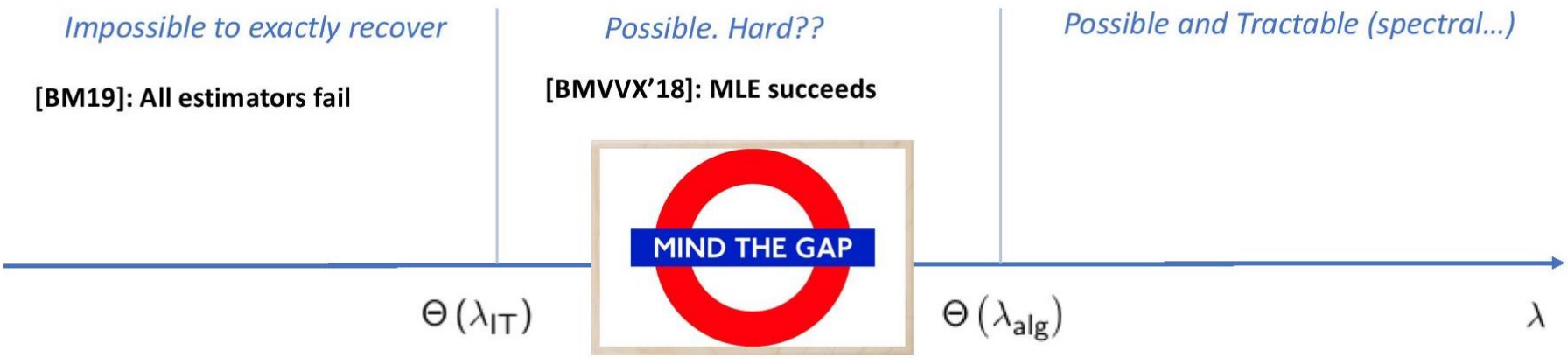


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$

- Generic in high dimensional inference: *computational-statistical gap*.

The Sparse PCA Model - Literature

Vast Literature. [Baik et al'05], [Johnstone, Lu '09], [Montanari et al '15], [Banks et al'18], [Ding et al'19], [Barbier et al'19], [Gamarnik et al '19]

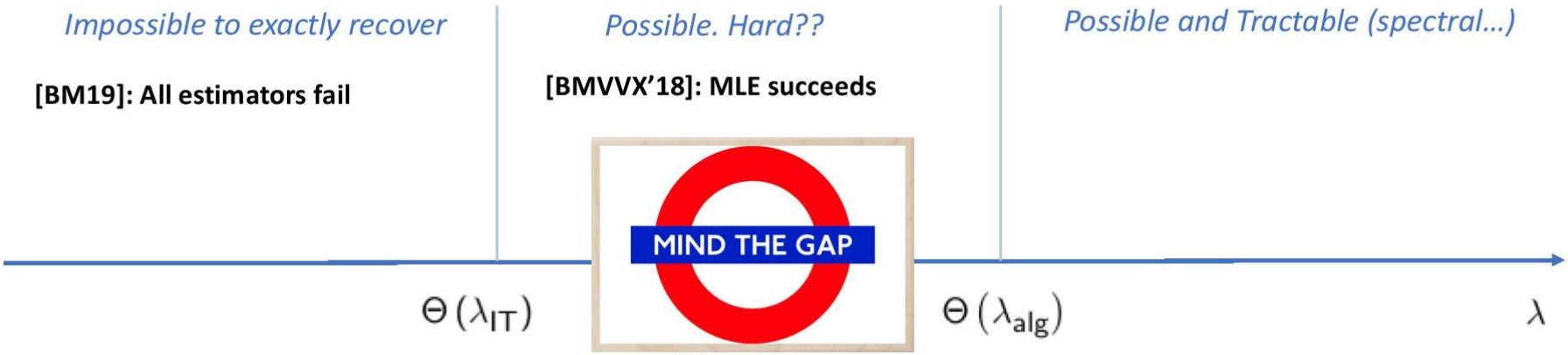


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$

- Generic in high dimensional inference: *computational-statistical gap*.
- *No* complexity-theoretic explanation.

The Sparse PCA Model - Literature

Vast Literature. [Baik et al'05], [Johnstone, Lu '09], [Montanari et al '15], [Banks et al'18], [Ding et al'19], [Barbier et al'19], [Gamarnik et al '19]

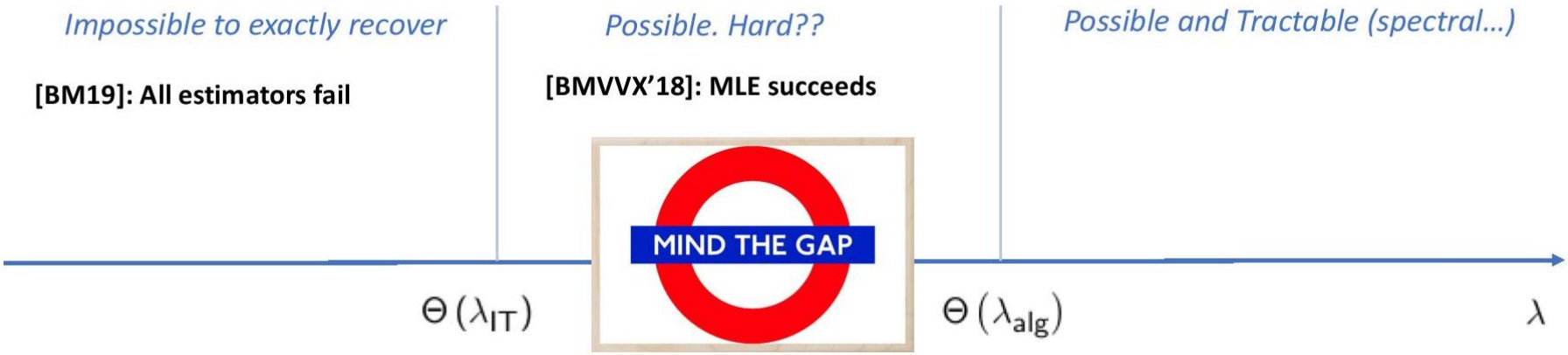


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$

- Generic in high dimensional inference: *computational-statistical gap*.
- *No* complexity-theoretic explanation.
- *Many methods:* low-degree (SOS), *statistical physics*, average-case reductions, statistical query lower bounds.

The Sparse PCA Model - Sub-exponential time methods

Recall $Y = \lambda xx^\top + W$.

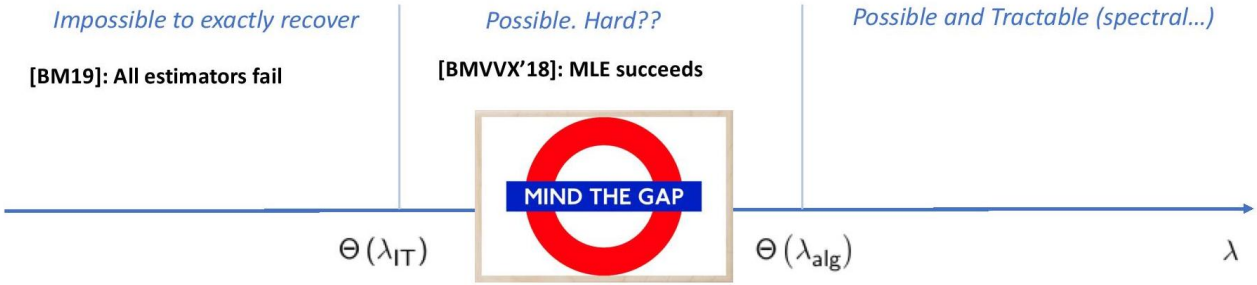


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$

The Sparse PCA Model - Sub-exponential time methods

Recall $Y = \lambda xx^\top + W$.

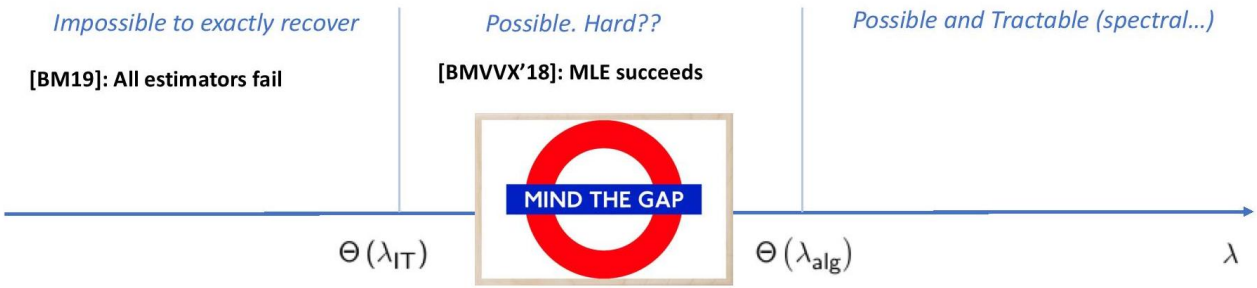


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Yv$

MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Yv$ can be solved in $e^{\tilde{\Theta}(k)}$-time. Optimal?

The Sparse PCA Model - Sub-exponential time methods

Recall $Y = \lambda xx^\top + W$.

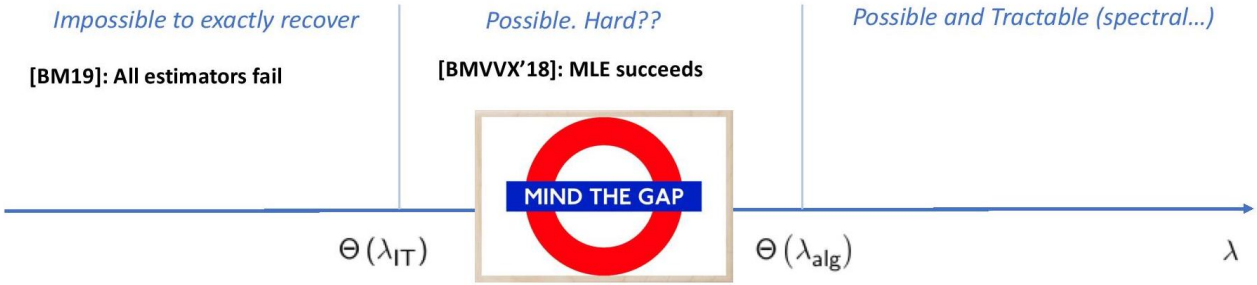


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Yv$

MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Yv$ can be solved in $e^{\tilde{\Theta}(k)}$-time. Optimal?

Conjecture [DKWB '19]-(low-degree method)

If $\lambda_{IT} < \lambda < \lambda_{alg}$, optimal algorithm takes $\exp\left(\tilde{\Theta}\left((\sqrt{n}\lambda)^{-2}\right)\right)$-time.

The Sparse PCA Model - Sub-exponential time methods

Recall $Y = \lambda xx^\top + W$.

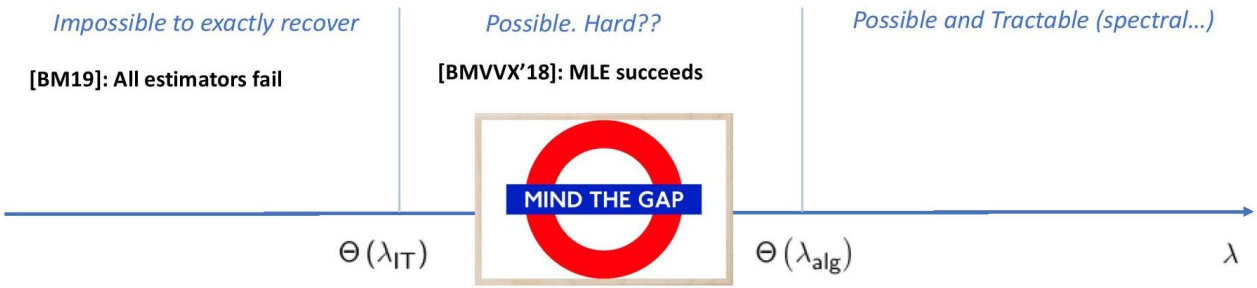


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Yv$

MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Yv$ can be solved in $e^{\tilde{\Theta}(k)}$-time. Optimal?

Conjecture [DKWB '19]-(low-degree method)

If $\lambda_{IT} < \lambda < \lambda_{alg}$, optimal algorithm takes $\exp\left(\tilde{\Theta}\left((\sqrt{n}\lambda)^{-2}\right)\right)$-time.

(Q1) MCMC methods? **(Q2)** Optimal time via "landscape" analysis ?

This Work

For **(almost) all** λ in the hard regime
we prove **lower bounds** for MCMC methods for solving the MLE
matching the conjectured $\exp\left(\tilde{\Omega}\left((\sqrt{n}\lambda)^{-2}\right)\right)$-time of [DKWB '19],
by establishing local barriers introduced in **spin glass theory**.

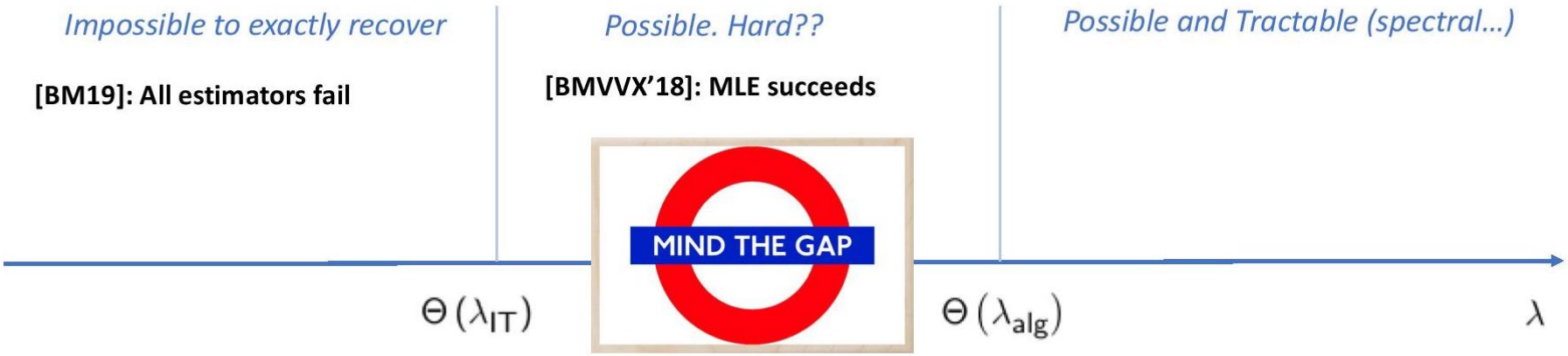


Figure: $\lambda_{IT} = 1/\sqrt{nk}$, $\lambda_{alg} = \min\{1/k, 1/\sqrt{n}\}$, MLE: $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$

Stationary distributions: Gibbs measures

Recall: $Y = \lambda xx^\top + W$
and MLE $\max_{v \in \{0,1\}^n, \|v\|_0 = k} v^\top Y v$ works for all $\lambda > \lambda_{IT}$.

Stationary distributions: Gibbs measures

Recall: $Y = \lambda xx^\top + W$
and MLE $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$ works for all $\lambda > \lambda_{IT}$.

- $\mathrm{MLE}_{k'}$: $\max_{v\in\{0,1\}^n:\|v\|_0=k'} v^\top Y v$.
- $k' \neq k$: optimal algorithm $k' = (\sqrt{n}\lambda)^{-2} < k$ [DKWB'19], and, (planted clique) $k' > k$ lifts local barriers (OGP) [G**Z**'19]

Stationary distributions: Gibbs measures

Recall: $Y = \lambda xx^\top + W$
and MLE $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$ works for all $\lambda > \lambda_{IT}$.

- $\mathrm{MLE}_{k'}$: $\max_{v\in\{0,1\}^n:\|v\|_0=k'} v^\top Y v$.
- $k' \neq k$: optimal algorithm $k' = (\sqrt{n}\lambda)^{-2} < k$ [DKWB'19], and, (planted clique) $k' > k$ lifts local barriers (OGP) [G**Z**'19]

The Gibbs measures

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}, \quad v \in \mathcal{B}_{k'} := \{v \in \{0,1\}^n, \|v\|_0 = k'\},$$

($\beta = \lambda n/2$ the posterior $\mathbb{P}[x|Y]$).

Stationary distributions: Gibbs measures

Recall: $Y = \lambda xx^\top + W$
and MLE $\max_{v\in\{0,1\}^n, \|v\|_0=k} v^\top Y v$ works for all $\lambda > \lambda_{IT}$.

- $MLE_{k'}$: $\max_{v\in\{0,1\}^n:\|v\|_0=k'} v^\top Y v$.
- $k' \neq k$: optimal algorithm $k' = (\sqrt{n}\lambda)^{-2} < k$ [DKWB'19], and, (planted clique) $k' > k$ lifts local barriers (OGP) [G**Z**'19]

The Gibbs measures

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}, \quad v \in \mathcal{B}_{k'} := \{v \in \{0,1\}^n, \|v\|_0 = k'\},$$

($\beta = \lambda n/2$ the posterior $\mathbb{P}[x|Y]$).

β, k' "informative"

Sampling v from $\mu_{\beta,k'}$ suffices for recovery,
i.e. $\langle v, x\rangle$ is "large" w.h.p.

Free Energy Wells (FEW) from Statistical Physics

The Gibbs measures

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}, v \in \mathcal{B}_{k'}.$$

Free Energy Wells (FEW) from Statistical Physics

The Gibbs measures

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}, v \in \mathcal{B}_{k'}.$$

For "overlap" $\ell = 0, 1, \ldots, \min\{k, k'\}$, $A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}$.

Free Energy Wells (FEW) from Statistical Physics

The Gibbs measures

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}, v \in \mathcal{B}_{k'}.$$

For "overlap" $\ell = 0, 1, \ldots, \min\{k, k'\}$, $A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}$.

We study the map $\ell \to \log \mu_{\beta,k'}(A_\ell)$ (essential for local dynamics).

Free Energy Wells (FEW) from Statistical Physics

The Gibbs measures

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}, v \in \mathcal{B}_{k'}.$$

For "overlap" $\ell = 0, 1, \ldots, \min\{k, k'\}$, $A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}$.

We study the map $\ell \to \log \mu_{\beta,k'}(A_\ell)$ (essential for local dynamics).

β, k' "informative" imply $\arg\max_\ell \mu_{\beta,k'}(A_\ell) \approx \min\{k', k\}$.

Free Energy Wells (FEW) from Statistical Physics

The Gibbs measures

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}, v \in \mathcal{B}_{k'}.$$

For "overlap" $\ell = 0, 1, \ldots, \min\{k, k'\}$, $A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}$.

We study the map $\ell \to \log \mu_{\beta,k'}(A_\ell)$ (essential for local dynamics).

β, k' "informative" imply $\arg\max_\ell \mu_{\beta,k'}(A_\ell) \approx \min\{k', k\}$.

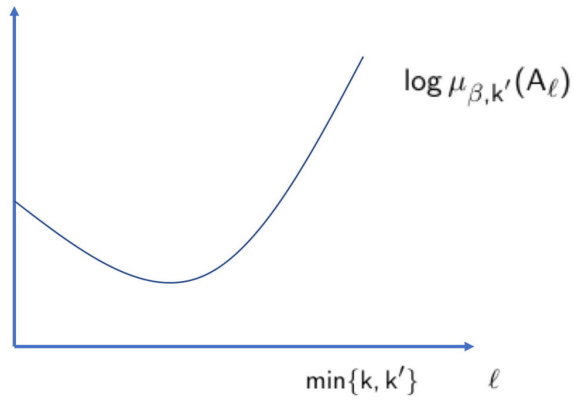


Figure: FEW

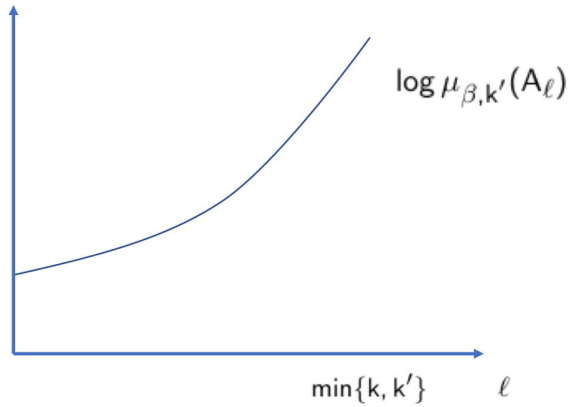


Figure: No FEW

Free Energy Wells (FEW) from Statistical Physics

The Gibbs measures

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}, v \in \mathcal{B}_{k'}.$$

For "overlap" $\ell = 0, 1, \ldots, \min\{k, k'\}$, $A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}$.

We study the map $\ell \to \log \mu_{\beta,k'}(A_\ell)$ (essential for local dynamics).

β, k' "informative" imply $\arg\max_\ell \mu_{\beta,k'}(A_\ell) \approx \min\{k', k\}$.

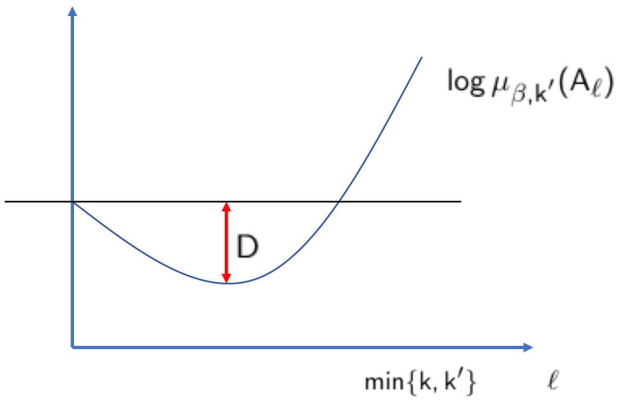


Figure: FEW of depth D

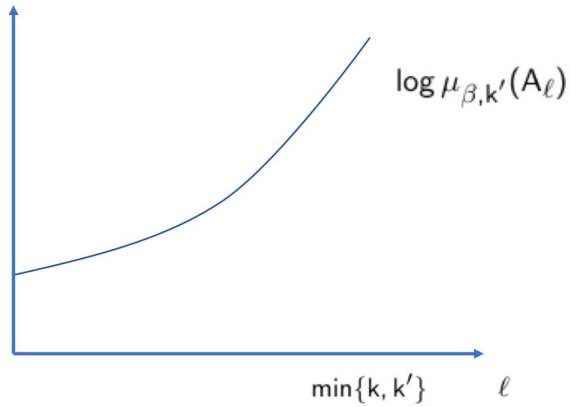


Figure: No FEW

A deep FEW slows MCMC methods

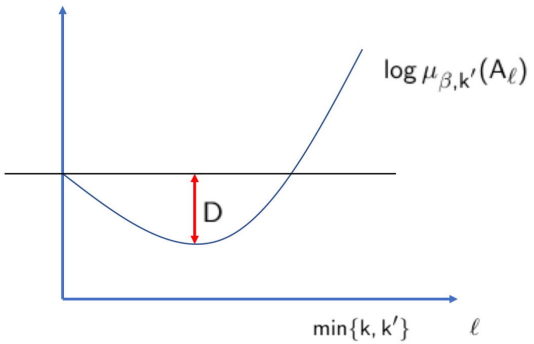


Figure: FEW of depth D

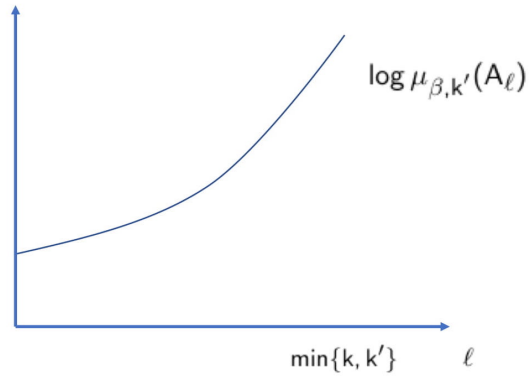


Figure: No FEW

A deep FEW slows MCMC methods

Theorem

Under FEW of depth D *for* $\mu_{\beta,k'}$, *any markov chain on* $\mathcal{B}_{k'}$ *which*

(1) *changes at most 2 coordinates at a step and*

(2) *has stationary distribution* $\mu_{\beta,k'}$,

requires (worst-case initialization) e^{D}*-time to recover* x.

Similar results for FEW in *tensor PCA* [BAGJ'18], *principal submatrix recovery* [GSJ'19] and *planted clique* [G**Z**'19].

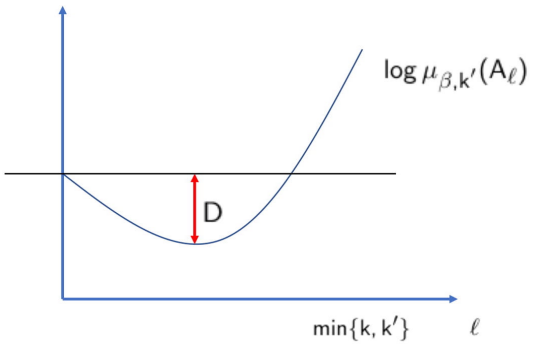


Figure: FEW of depth D

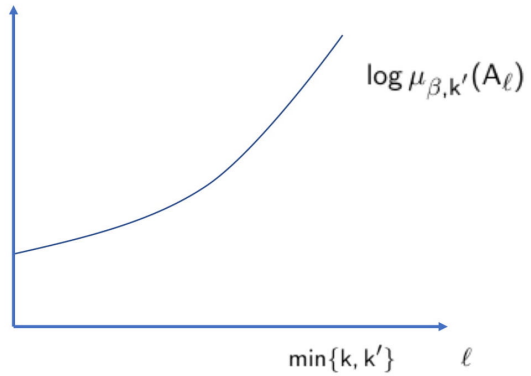


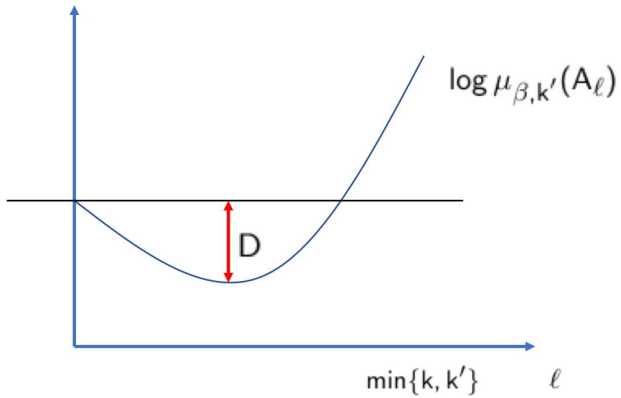
Figure: No FEW

Main Result on FEW

Theorem

For (almost) all λ in the "hard" regime and all β, k' "informative", $\mu_{\beta,k'}$ admits a FEW of depth $\Omega\left((\sqrt{n}\lambda)^{-2}\right)$.*

In particular, any "local" MCMC method requires $e^{\Omega\left((\sqrt{n}\lambda)^{-2}\right)}$-time.



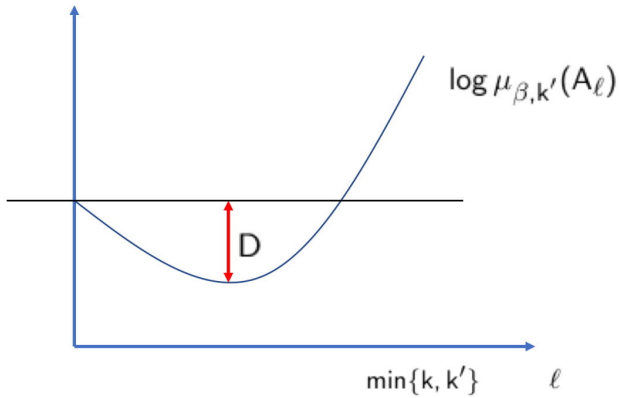
Main Result on FEW

Theorem

For (almost) all λ in the "hard" regime and all β, k' "informative", $\mu_{\beta,k'}$ admits a FEW of depth $\Omega\left((\sqrt{n}\lambda)^{-2}\right)$.*

In particular, any "local" MCMC method requires $e^{\Omega\left((\sqrt{n}\lambda)^{-2}\right)}$-time.

- **Proves the conjecture** of [DKWB'19] for MCMC methods! Optimal $\max_{v\in\{0,1\}^p, \|v\|_0=k'} v^\top Y v, k' \approx (\sqrt{n}\lambda)^{-2}$ cannot be boosted by MCMC.



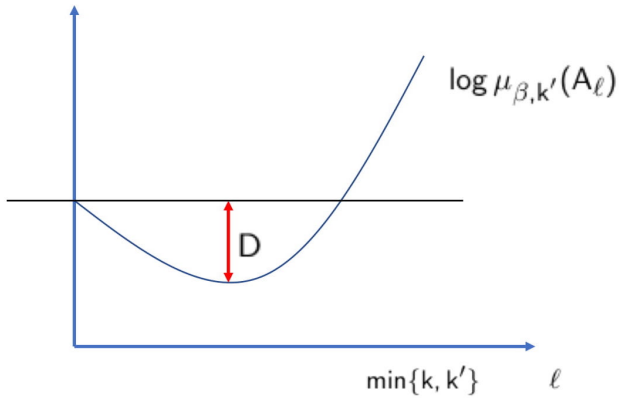
Main Result on FEW

Theorem

For (almost) all λ in the "hard" regime and all β, k' "informative", $\mu_{\beta,k'}$ admits a FEW of depth $\Omega\left((\sqrt{n}\lambda)^{-2}\right)$.*

In particular, any "local" MCMC method requires $e^{\Omega\left((\sqrt{n}\lambda)^{-2}\right)}$-time.

- **Proves the conjecture** of [DKWB'19] for MCMC methods! Optimal $\max_{v\in\{0,1\}^p, \|v\|_0=k'} v^\top Y v, k' \approx (\sqrt{n}\lambda)^{-2}$ cannot be boosted by MCMC.
- **Almost all*** becomes **all** under $k \le n^{1/3}$ or $k' \le k$.



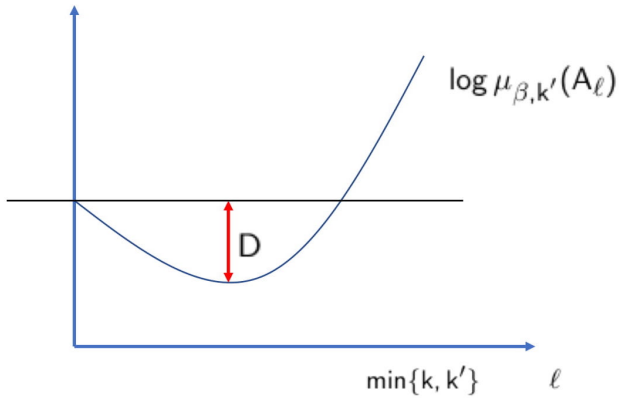
Main Result on FEW

Theorem

For (almost) all λ in the "hard" regime and all β, k' "informative", $\mu_{\beta,k'}$ admits a FEW of depth $\Omega\left((\sqrt{n}\lambda)^{-2}\right)$.*

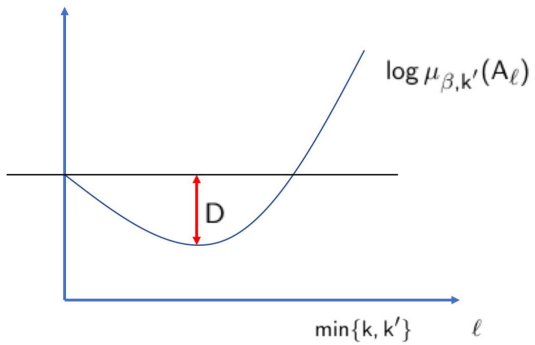
In particular, any "local" MCMC method requires $e^{\Omega\left((\sqrt{n}\lambda)^{-2}\right)}$-time.

- **Proves the conjecture** of [DKWB'19] for MCMC methods! Optimal $\max_{v\in\{0,1\}^p, \|v\|_0=k'} v^\top Y v, k' \approx (\sqrt{n}\lambda)^{-2}$ cannot be boosted by MCMC.
- **Almost all*** becomes **all** under $k \le n^{1/3}$ or $k' \le k$.
- Establish **OGP** ($\beta = \infty$) for the same parameters.



Proof Sketch

$$\mu_{\beta,k'}(A_\ell) \propto \sum_{v \in A_\ell} e^{\beta v^\top Y v}, \quad A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}.$$



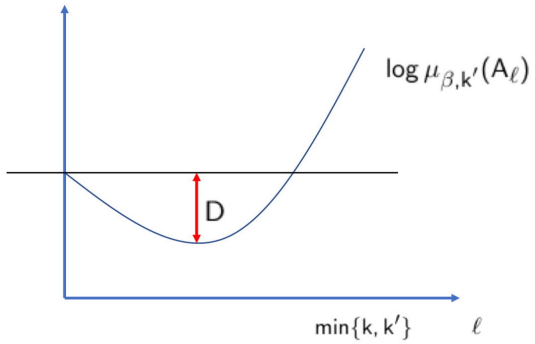
Proof Sketch

$$\mu_{\beta,k'}(A_\ell) \propto \sum_{v \in A_\ell} e^{\beta v^\top Y v}, \quad A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}.$$

- "Small" $\beta < \beta_1$: $\mu_{\beta,k'}$ like uniform, so for "small" ℓ,

$$\log \mu_{\beta,k'}(A_\ell) \approx \log |A_\ell| - \text{const.}$$

Entropy argument gives depth D "roughly" $\Omega(1/(\beta\lambda))$



Proof Sketch

$$\mu_{\beta,k'}(A_\ell) \propto \sum_{v \in A_\ell} e^{\beta v^\top Y v}, \quad A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}.$$

- "Small" $\beta < \beta_1$: $\mu_{\beta,k'}$ like uniform, so for "small" ℓ,

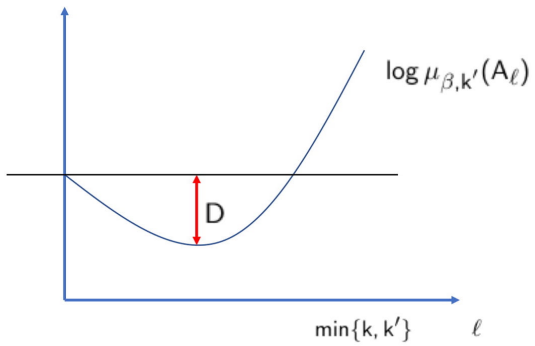
$$\log \mu_{\beta,k'}(A_\ell) \approx \log |A_\ell| - \text{const.}$$

Entropy argument gives depth D "roughly" $\Omega\left(1/(\beta\lambda)\right)$

- "Large" $\beta > \beta_2$ (but $\beta_2 < \beta_1$!):

$$\log \mu_{\beta,k'}(A_\ell) \approx \beta \max_{v \in A_\ell} v^\top Y v.$$

2nd moment method gives depth $D = \Omega\left(k'\right)$.



Proof Sketch

$$\mu_{\beta,k'}(A_\ell) \propto \sum_{v \in A_\ell} e^{\beta v^\top Y v}, \quad A_\ell := \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}.$$

- "Small" $\beta < \beta_1$: $\mu_{\beta,k'}$ like uniform, so for "small" ℓ,

$$\log \mu_{\beta,k'}(A_\ell) \approx \log |A_\ell| - \text{const.}$$

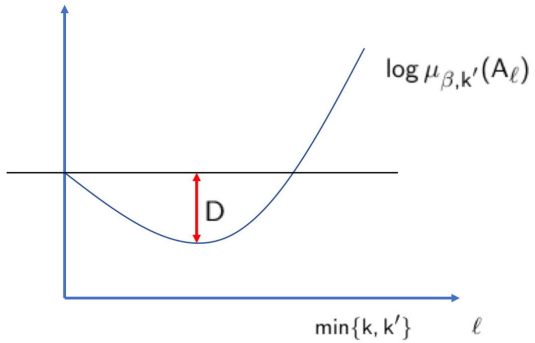
Entropy argument gives depth D "roughly" $\Omega\left(1/(\beta\lambda)\right)$

- "Large" $\beta > \beta_2$ (but $\beta_2 < \beta_1$!):

$$\log \mu_{\beta,k'}(A_\ell) \approx \beta \max_{v \in A_\ell} v^\top Y v.$$

2nd moment method gives depth $D = \Omega\left(k'\right)$.

- Criticality at $\beta = \beta_{\text{Bayes}} = \lambda n/2$ gives depth $\Omega\left((\sqrt{n}\lambda)^{-2}\right)$.



Summary/Future Work

- **Optimal sub-exponential MCMC lower bounds** for Sparse PCA: using the depth of FEW from stat physics.

Summary/Future Work

- **Optimal sub-exponential MCMC lower bounds** for Sparse PCA: using the depth of FEW from stat physics.
- **Same with low-degree methods** prediction, but using the landscape! How general is the connection?

Summary/Future Work

- **Optimal sub-exponential MCMC lower bounds** for Sparse PCA: using the depth of FEW from stat physics.
- **Same with low-degree methods** prediction, but using the landscape! How general is the connection?
- **Big open problem:** Positive MCMC results for inference. Jerrum's ['92] planted clique (positive) question open!

Summary/Future Work

- **Optimal sub-exponential MCMC lower bounds** for Sparse PCA: using the depth of FEW from stat physics.
- **Same with low-degree methods** prediction, but using the landscape! How general is the connection?
- **Big open problem:** Positive MCMC results for inference. Jerrum's ['92] planted clique (positive) question open!

Thank you!!

$$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v},\ \ell \to \log \mu_{\beta,k'}(A_\ell),\ A_\ell = \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}.$$

$\mu_{\beta,k'}(v) \propto e^{\beta v^\top Y v}$, $\ell \to \log \mu_{\beta,k'}(A_\ell)$, $A_\ell = \{v \in \mathcal{B}_{k'} : \langle v, x \rangle = \ell\}$.

- If $\beta \approx \infty$,

$$\log \mu_{\beta,k'}(A_\ell) \approx \beta \max_{v \in A_\ell} v^\top Y v.$$

Non-monotonicity iff OGP.

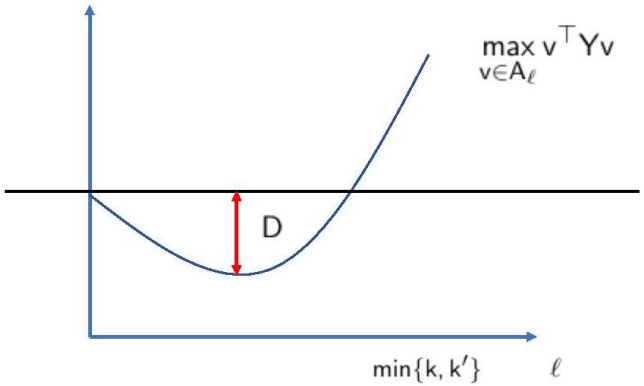


Figure: $\beta = \infty$, OGP of depth D

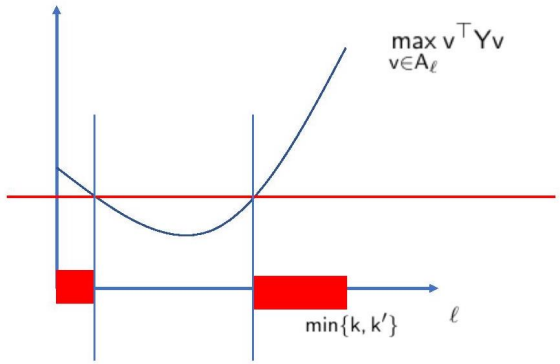


Figure: $\beta = \infty$, OGP