Policy Gradient mathematics

A/Prof Richard Yi Da Xu
richardxu.com

University of Technology Sydney (UTS)

July 23, 2019

Content

1. Policy Gradient Theorem
2. Mathematics on Trusted Region Optimization in RL
3. Natural Gradients on TRPO
4. Proximal Policy Optimization (PPO)
5. Conjugate Gradient Algorithm

This lecture is referenced heavily from:

- https://lilianweng.github.io/lil-log/2018/04/08/policy-gradient-algorithms.html. I borrowed it heavily, please check her goodies on RL and GAN
- https://medium.com/@jonathan_hui/rl-trust-region-policy-optimization-trpo-explained-a6ee04eeeee9, Jonathan Hui's blog. Again, lots of goodies.
- http://www.cs.cmu.edu/~pradeepr/convexopt/Lecture_Slides/conjugate_direction_methods.pdf

What is Policy Gradient "on the surface"

- Gradient of Expected entire Rewards $R(\tau)$ collected by taking a "trajectory" τ following π_θ:

$$
\begin{aligned}
\nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] &= \mathbb{E}_{\tau\sim\pi_\theta}\left[R(\tau)\cdot \nabla_\theta \log \mathbb{P}_\theta(\tau)\right] \\
&= \mathbb{E}_{\tau\sim\pi_\theta}\left[R(\tau)\cdot \nabla_\theta \left(\sum_{t=0}^{T-1} \log \pi_\theta(a_t|s_t)\right)\right]
\end{aligned}
$$

- Derivative of Log-likelihood of Policy Gradient is:

$$
\begin{aligned}
\nabla_\theta \log \mathbb{P}_\theta(\tau) &= \nabla_\theta \log\left(\mu(s_0)\prod_{t=0}^{T-1}\pi_\theta(a_t|s_t)P(s_{t+1}|s_t,a_t)\right) \\
&= \nabla_\theta\left[\underbrace{\log \mu(s_0)}_{\text{no }\theta} + \sum_{t=0}^{T-1}\left(\log \pi_\theta(a_t|s_t) + \underbrace{\log P(s_{t+1}|s_t,a_t)}_{\text{no }\theta}\right)\right] \\
&= \nabla_\theta \sum_{t=0}^{T-1}\log \pi_\theta(a_t|s_t)
\end{aligned}
$$

- $\log p(s_{t+1}|s_t,a_t)$ has no θ is controversial, we need to see why

Significance of Policy Gradient Theorem

- we use an alternative representation:

$$J(\theta) \equiv V^{\pi}(s_0)$$

  which we can expand using recursion as needed for unknow T:
- Computing gradient $\nabla_\theta J(\theta)$ is **difficult** because it depends on both:
  1. action selection **directly** determined by π_θ, and
  2. stationary state following action selection behavior **indirectly** determined by π_θ
- difficult to estimate policy update effect on state distribution for generally unknown environment
- however, **Policy gradient theorem** states:

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} Q^{\pi}(s, a)\pi_\theta(a|s) \\ &\propto \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} Q^{\pi}(s, a)\nabla_\theta \pi_\theta(a|s)\end{aligned}$$

- **significance**: above objective function does **not** involve derivative of state distribution $d^{\pi}(.)$

Proof of Policy Gradient Theorem

- We want a policy to maximize $J(\theta) \equiv V^{\pi}(s)$:
- first step is always to write $V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi_{\theta}(a|s) Q^{\pi}(s,a)$:

$$
\begin{aligned}
\nabla_{\theta} V^{\pi}(s) &= \nabla_{\theta} \Big(\sum_{a \in \mathcal{A}} \underbrace{\pi_{\theta}(a|s)}_{u} \underbrace{Q^{\pi}(s,a)}_{v} \Big) \\
&= \sum_{a \in \mathcal{A}} \Big(\underbrace{\nabla_{\theta} \pi_{\theta}(a|s) Q^{\pi}(s,a)}_{\phi(s)} + \pi_{\theta}(a|s) \nabla_{\theta} Q^{\pi}(s,a) \Big) \\
&= \sum_{a \in \mathcal{A}} \Big(\phi(s) + \pi_{\theta}(a|s) \nabla_{\theta} \sum_{s'} \sum_{r} P(s', r|s,a) (\underbrace{r + V^{\pi}(s')}_{\text{immediate \& future reward}}) \Big) \\
&= \sum_{a \in \mathcal{A}} \Big(\phi(s) + \pi_{\theta}(a|s) \sum_{s'} \sum_{r} P(s', r|s,a) \nabla_{\theta} V^{\pi}(s') \Big) \qquad \text{remove part independent of } \theta \\
&= \sum_{a \in \mathcal{A}} \Big(\phi(s) + \pi_{\theta}(a|s) \sum_{s'} P(s'|s,a) \nabla_{\theta} V^{\pi}(s') \Big) \qquad \text{retain marginal by integrate out } r
\end{aligned}
$$

Policy Gradient Theorem (1)

- let $\rho^{\pi}(s \to s', t)$ to be the probability of transition from state $s \to s'$ in t steps.

$$
\begin{aligned}
\nabla_\theta V^{\pi}(s) =& \phi(s) + \sum_a \pi_\theta(a|s) \sum_{s'} P(s'|s,a) \nabla_\theta V^{\pi}(s') \\
=& \phi(s) + \sum_{s'} \underline{\sum_a \pi_\theta(a|s) P(s'|s,a)} \nabla_\theta V^{\pi}(s') && \text{switch the two summation places} \\
=& \phi(s) + \sum_{s'} \underline{\rho^{\pi}(s \to s', 1)} \nabla_\theta V^{\pi}(s') && \text{expand this recursion: } s \to s' \text{ and } s' \to s'' \\
=& \phi(s) + \sum_{s'} \rho^{\pi}(s \to s', 1) \left[\phi(s') + \sum_{s''} \rho^{\pi}(s' \to s'', 1) \nabla_\theta V^{\pi}(s'') \right] \\
=& \phi(s) + \sum_{s'} \rho^{\pi}(s \to s', 1) \phi(s') + \underbrace{\sum_{s'} \sum_{s''} \rho^{\pi}(s \to s', 1) \rho^{\pi}(s' \to s'', 1)}\ \nabla_\theta V^{\pi}(s'') \\
=& \phi(s) + \sum_{s'} \rho^{\pi}(s \to s', 1) \phi(s') + \underbrace{\sum_{s''} \rho^{\pi}(s \to s'', 2)} \quad \underbrace{\nabla_\theta V^{\pi}(s'')} && \text{Repeatedly expand } \nabla_\theta V^{\pi}(.)\text{:} \\
=& \underbrace{\rho^{\pi}(s \to s, 0)}_{=1} \phi(s) + \sum_{s^{(1)} \in \mathcal{S}} \rho^{\pi}(s \to s^{(1)}, 1) \phi(s^{(1)}) + \sum_{s^{(2)} \in \mathcal{S}} \rho^{\pi}(s \to s^{(2)}, 2) \phi(s^{(2)}) + \ldots \\
=& \sum_{\{s^{(t)}\} \in \mathcal{S}} \sum_{t=0}^{\infty} \rho^{\pi}(s \to s^{(t)}, t) \phi(s^{(t)})
\end{aligned}
$$

Policy Gradient Theorem (2)

- starting from a state s_0:

$$\nabla_\theta J(\theta) \equiv \nabla_\theta V^\pi(s_0)$$

$$= \sum_s \underbrace{\sum_{t=0}^{\infty} \rho^\pi(s_0 \to s, t)}_{\eta(s)} \phi(s)$$

$$= \sum_s d^\pi(s) \sum_a \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) \quad \text{where } d^\pi(s) \equiv \frac{\eta(s)}{\sum_s \eta(s)} \text{ is a normalized version of } \eta(s)$$

$d^\pi(s)$ acts like the weight of derivative concerning particular s

- to write $\nabla_\theta J(\theta)$ in terms of $\mathbb{E}_\pi[.]$

$$\nabla_\theta J(\theta) \propto \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} Q^\pi(s,a) \nabla_\theta \pi_\theta(a|s)$$

$$= \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \pi_\theta(a|s) Q^\pi(s,a) \underbrace{\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}}$$

$$= \underbrace{\sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \pi_\theta(a|s)}_{\mathbb{E}_\pi} \left[Q^\pi(s,a) \underbrace{\nabla_\theta \log \pi_\theta(a|s)} \right]$$

$$= \mathbb{E}_\pi \left[Q^\pi(s,a) \nabla_\theta \ln \pi_\theta(a|s) \right]$$

Variance reduction using Baseline

- subtract a baseline function $B(s)$ from policy gradient, note $B(s)$ only depends on state s, **not depends on action** a, such that:

$$\mathbb{E}_\pi \Big[\underbrace{Q^\pi(s, a)}_{\text{replace with} B(s)} \nabla_\theta \ln \pi_\theta(a|s) \Big]$$

$$\text{so we have: } \mathbb{E}_\pi \left[B(s) \nabla_\theta \ln \pi_\theta(a|s) \right]$$

$$= \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(s, a) B(s)$$

$$= \sum_{s \in \mathcal{S}} d^\pi(s) B(s) \nabla_\theta \sum_{a \in \mathcal{A}} \pi_\theta(s, a)$$

$$= 0$$

- A good baseline is $B(s) = V^\pi(s)$:

$$\begin{aligned}
\text{without baseline} \qquad & \nabla_\theta J(\theta) = \mathbb{E}_\pi \left[Q^\pi(s, a) \nabla_\theta \ln \pi_\theta(a|s) \right] \\
\text{with baseline} \qquad & \nabla_\theta J(\theta) = \mathbb{E}_\pi \left[\nabla_\theta \ln \pi_\theta(a|s) (Q^\pi(s, a) - V^\pi(s)) \right] \\
& \qquad\quad = \mathbb{E}_\pi \left[\nabla_\theta \ln \pi_\theta(a|s) A^\pi(s, a) \right]
\end{aligned}$$

Off policy

- change behavioral distribution from π to β but target policy is still $\pi_\theta(a|s)$:

$$J(\theta) = \sum_{s \in \mathcal{S}} d^{\beta}(s) \sum_{a \in \mathcal{A}} Q^{\pi}(s,a)\pi_\theta(a|s) = \mathbb{E}_{s \sim d^{\beta}}\Big[\sum_{a \in \mathcal{A}} Q^{\pi}(s,a)\pi_\theta(a|s)\Big]$$

- adding Importance sampling

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \mathbb{E}_{s \sim d^{\beta}}\Big[\sum_{a \in \mathcal{A}} \underbrace{Q^{\pi}(s,a)}_{u}\,\underbrace{\pi_\theta(a|s)}_{v}\Big] \\
&= \mathbb{E}_{s \sim d^{\beta}}\Big[\sum_{a \in \mathcal{A}} \big(Q^{\pi}(s,a)\nabla_\theta \pi_\theta(a|s) + \pi_\theta(a|s)\nabla_\theta Q^{\pi}(s,a)\big)\Big] \\
&\overset{(i)}{\approx} \mathbb{E}_{s \sim d^{\beta}}\Big[\sum_{a \in \mathcal{A}} Q^{\pi}(s,a)\nabla_\theta \pi_\theta(a|s)\Big] \qquad \text{big assumption: Ignore the red part:} \\
&= \mathbb{E}_{s \sim d^{\beta}}\Big[\sum_{a \in \mathcal{A}} \beta(a|s)\frac{\pi_\theta(a|s)}{\beta(a|s)} Q^{\pi}(s,a)\frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)}\Big] \\
&= \mathbb{E}_{\beta}\Big[\underbrace{\frac{\pi_\theta(a|s)}{\beta(a|s)}}_{\text{importance weights}} Q^{\pi}(s,a)\nabla_\theta \ln \pi_\theta(a|s)\Big]
\end{aligned}$$

- using $\beta = \pi_{k\theta_k}(a|s)$, you have on-policy, so we use β genericaly

Trust region policy optimization (TRPO)

- look at the equation for off-policy + baseline:

$$\nabla_\theta J(\theta) = \mathbb{E}_\beta \Big[\underbrace{\frac{\pi_\theta(a|s)}{\beta(a|s)}} \nabla_\theta \ln \pi_\theta(a|s) \big(Q^\pi(s,a) - V^\pi(s) \big) \Big]$$

- θ_k is the policy before update, as we do not need to update each time. It can be made same as β (then we have on-policy)

$$\begin{aligned} J(\theta) &= \sum_{s \in \mathcal{S}} \rho^{\pi_{\theta_{\text{old}}}} \sum_{a \in \mathcal{A}} \left(\pi_\theta(a|s) \hat{A}_{\theta_{\text{old}}}(s,a) \right) \\ &= \sum_{s \in \mathcal{S}} \rho^{\pi_{\theta_{\text{old}}}} \sum_{a \in \mathcal{A}} \left(\beta(a|s) \frac{\pi_\theta(a|s)}{\beta(a|s)} \hat{A}_{\theta_{\text{old}}}(s,a) \right) \\ &= \mathbb{E}_{s \sim \rho^{\pi_{\theta_{\text{old}}}}, a \sim \beta} \left[\frac{\pi_\theta(a|s)}{\beta(a|s)} \hat{A}_{\theta_{\text{old}}}(s,a) \right] \end{aligned}$$

- as a side note, if we were to take derivatives to compute for gradient descent:

$$\begin{aligned} \nabla_\theta J(\theta) &= \sum_{s \in \mathcal{S}} \rho^{\pi_{\theta_{\text{old}}}} \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) \hat{A}_{\theta_{\text{old}}}(s,a) \\ &= \sum_{s \in \mathcal{S}} \rho^{\pi_{\theta_{\text{old}}}} \sum_{a \in \mathcal{A}} \beta(a|s) \frac{\pi_\theta(a|s)}{\beta(a|s)} \frac{\nabla_\theta \pi_\theta(a|s)}{\pi_\theta(a|s)} \hat{A}_{\theta_{\text{old}}}(s,a) \\ &= \mathbb{E}_{s \sim \rho^{\pi_{\theta_{\text{old}}}}, a \sim \beta} \left[\frac{\pi_\theta(a|s)}{\beta(a|s)} \log \left(\nabla_\theta \pi_\theta(a|s) \right) \hat{A}_{\theta_{\text{old}}}(s,a) \right] \end{aligned}$$

TRPO objective

- objective function, assume we let $\beta \equiv \theta_{\text{old}}$:

$$\max_{\pi} \left(J(\pi) - J(\beta) \right)$$

- basically, finding the best new policy π to improve upon the previous behavioral policy β
- however, we need it to:

$$\begin{aligned}
&\max_{\pi}(J(\pi) - J(\beta)) \\
&\qquad J(\pi) - J(\beta) \geq \mathcal{L}_{\beta}(\pi) - C\, \mathbb{E}_{s \sim d_k^{\beta}}[\mathrm{KL}(\pi \| \beta)(s)] \\
&\qquad\qquad = \underbrace{\mathbb{E}_{\tau \sim \beta}\left[\sum_{t=0}^{\infty} \gamma^t \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t) \right] - C\, \mathbb{E}_{s \sim d_k^{\beta}}[\mathrm{KL}(\pi \| \beta)[s]]}_{\text{lower bound } \mathcal{L}_{\beta}(\pi)}
\end{aligned}$$

- so we just need to maximize $\mathcal{L}_{\beta}(\pi)$ instead

Why equality occurs ($\pi = \beta$)?

$$J(\beta) - J(\beta) = \underbrace{\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t \frac{\beta(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t)\right]}_{=\text{what?}} - \underbrace{C\,\mathbb{E}_{s\sim d_k^{\beta}}[\text{KL}(\beta\,\|\,\beta)[s]]}_{=0,\text{ well, that's KL}}$$

- looking at $\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t A^{\beta}(s_t, a_t)\right]$:

$$\begin{aligned}
\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t A^{\beta}(s_t, a_t)\right] &= \sum_{t=0}^{\infty}\gamma^t \sum_{a_t\in\mathcal{A}} A^{\beta}(s_t, a_t)\\
&= \sum_{t=0}^{\infty}\gamma^t \sum_{a_t\in\mathcal{A}} \Big(Q^{\beta}(s_t, a_t) - V^{\beta}(s_t)\Big)\\
&= \sum_{t=0}^{\infty}\gamma^t \Big(\sum_{a_t\in\mathcal{A}} Q^{\beta}(s_t, a_t)\Big) - V^{\beta}(s_t)\\
&= \sum_{t=0}^{\infty}\gamma^t \Big(V^{\beta}(s_t) - V^{\beta}(s_t)\Big) = 0
\end{aligned}$$

- As a side note: if instead we look at $\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t f(a_t) A^{\beta}(s_t, a_t)\right]$:

$$\begin{aligned}
\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty} f(a_t)\gamma^t A^{\beta}(s_t, a_t)\right] &= \sum_{t=0}^{\infty}\gamma^t \sum_{a_t\in\mathcal{A}} f(a_t)\Big(Q^{\beta}(s_t, a_t) - V^{\beta}(s_t)\Big)\\
&= \sum_{t=0}^{\infty}\gamma^t \underbrace{\Big(\sum_{a_t\in\mathcal{A}} f(a_t) Q^{\beta}(s_t, a_t)\Big)}_{\neq f(a_t)V^{\beta}(s_t)} - f(a_t)V^{\beta}(s_t)
\end{aligned}$$

Always improve the result

- we know,

$$
\begin{aligned}
& J(\beta) - J(\beta) = 0, \text{ and, } J(\pi) - J(\beta) \geq \mathcal{L}_\beta(\pi) \\
\implies & J(\pi) - J(\beta) \geq 0 \text{ after we optimized } \mathcal{L}_\beta(\pi)
\end{aligned}
$$

- meaning the new policy is always as good as the previous one

KL penalized vs KL constrained

Two different constraints for $\text{KL}(\pi\|\beta)$

- $\text{KL}(\pi\|\beta) = C$:

$$\max_{\pi}\left[\underbrace{\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t\frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}A^{\beta}(s_t,a_t)\right]}_{\mathcal{L}_{\theta_k}(\theta)} - C\ \text{KL}(\pi\|\beta)\right]$$

- $\text{KL}(\pi\|\beta) \leq \delta$:

$$\max_{\pi}\left[\underbrace{\mathbb{E}_{\tau\sim\beta}\left[\sum_{t=0}^{\infty}\gamma^t\frac{\pi(a_t|s_t)}{\beta(a_t|s_t)}A^{\beta}(s_t,a_t)\right]}_{\mathcal{L}_{\theta_k}(\theta)}\right]$$

$$\text{s.t.}\quad \text{KL}(\pi\|\beta) \leq \delta$$

- solving the above is hard, we approx both $\mathcal{L}_{\theta_k}(\theta)$ and $\text{KL}(\pi\|\beta)$ part
- $\text{KL}(\pi\|\beta)$ part need concept of **natural gradient**

Natural Gradient manifold: KL$(\pi \| \beta) = C$

- Taylor (order 1) expansion of $\mathcal{L}(\theta)$:

$$\mathcal{L}(\theta + h) \approx \mathcal{L}(\theta) + \nabla_\theta \mathcal{L}(\theta)^\top h$$
$$\implies \arg\min_h \{\mathcal{L}(\theta + h)\} \approx \arg\min_h \{\nabla_\theta \mathcal{L}(\theta)^\top h\}$$

- look at steepest gradient descent: we minimize at an **equiv-euclidean-distance** hyper-sphere:

$$\begin{aligned} h^* &= \arg\min_h \{\mathcal{L}(\theta + h) : \|h\| = 1\} \\ &\approx \arg\min_h \{\nabla_\theta \mathcal{L}(\theta)^\top h : \|h\| = 1\} \\ &= -\nabla_\theta \mathcal{L}(\theta) \end{aligned}$$

- now instead, we minimize at an **equiv-KL-distance** manifold:

$$\begin{aligned} h^* &= \arg\min_h \left\{\mathcal{L}(\theta + h) : h \in \left(\mathrm{KL}[p_\theta \| p_{\theta+h}] = c\right)\right\} \\ &\approx \arg\min_h \left\{\nabla_\theta \mathcal{L}(\theta)^\top h : h \in \left(\mathrm{KL}[p_\theta \| p_{\theta+h}] = c\right)\right\} \end{aligned}$$

Natural Gradient manifold: $\text{KL}(\pi \| \beta) = C$

- solving

$$h^* \approx \arg\min_h \left\{ \nabla_\theta \mathcal{L}(\theta)^\top h : h \in \left(\text{KL}[p_\theta \| p_{\theta+h}] = c \right) \right\}$$

- solve using Lagrange Multiplier:

$$= \arg\min_h \left(\nabla_\theta \mathcal{L}(\theta)^\top h + \lambda (\text{KL}[p_\theta \| p_{\theta+h}] - c) \right)$$

- **if** we can prove second degree Taylor approximation:

$$\text{KL}[p_\theta \| p_{\theta+h}] \equiv \text{KL}[p(x|\theta) \| p(x|\theta + h)] \approx \frac{1}{2} h^\top \text{F} h \qquad \text{A}$$

- then,

$$\begin{aligned} h^* &\approx \arg\min_h \left(\nabla_\theta \mathcal{L}(\theta)^\top h + \lambda \Big(\frac{1}{2} h^\top \text{F} h - c \Big) \right) \\ \implies & \frac{\partial}{\partial h} \left(\nabla_\theta \mathcal{L}(\theta)^\top h + \frac{1}{2} \lambda\, h^\top \text{F} h - \lambda c \right) = 0 \\ & \nabla_\theta \mathcal{L}(\theta) + \lambda\, \text{F} h = 0 \\ & h = -\frac{1}{\lambda} \text{F}^{-1} \nabla_\theta \mathcal{L}(\theta) \end{aligned}$$

- look at taylor expansion:

$$f(x_0 + h) \approx f(\mathbf{x}) + f'(\mathbf{x})h + \frac{1}{2} h^\top f''(\mathbf{x})h \quad | \ \mathbf{x} = x_0$$

- to avoide confusion: $x_0 \to \theta_0$ is constant, and $\theta' \to \theta$ is variable

$$\begin{aligned}
\text{KL}[p_{\theta_0} \parallel p_{\theta+h}] &\approx \text{KL}[p_{\theta_0} \parallel p_\theta] + \left((\nabla_\theta \text{KL}[p_{\theta_0} \parallel p_\theta])^\top h + \frac{1}{2} h^\top (\nabla^2_\theta \text{KL}[p_{\theta_0} \parallel p_\theta]) h \right)\Big|_{\theta=\theta_0} \\
&= \text{KL}[p_{\theta_0} \parallel p_{\theta_0}] + \underbrace{\left(\nabla_\theta \text{KL}[p_{\theta_0} \parallel p_\theta]\Big|_{\theta=\theta_0}\right)^\top}_{1} h + \frac{1}{2} h^\top \underbrace{\left(\nabla^2_\theta \text{KL}[p_{\theta_0} \parallel p_\theta]\Big|_{\theta=\theta_0}\right)}_{2} h \\
&= 0 + 0 + \frac{1}{2} h^\top \text{F} h \\
&= \frac{1}{2} h^\top \text{F} h
\end{aligned}$$

- note the ordering when computing $\nabla_\theta f(\theta, \theta_0)\Big|_{\theta=\theta_0}$: take derivative first, then substitute.
- look at KL between $p(x|\theta)$ and $p(x|\theta')$:

$$\text{KL}[p(x|\theta) \parallel p(x|\theta')] = \mathbb{E}_{p(x|\theta)}\left[\log \frac{p(x|\theta)}{p(x|\theta')}\right] = \mathbb{E}_{p(x|\theta)}[\log p(x|\theta)] - \mathbb{E}_{p(x|\theta)}[\log p(x|\theta')]$$

- taking first derivative with respect to θ':

$$\begin{aligned}
\nabla_{\theta'}\text{KL}[p(x|\theta) \parallel p(x|\theta')] &= \nabla_{\theta'}\left[\mathbb{E}_{p(x|\theta)}[\log p(x|\theta)] - \mathbb{E}_{p(x|\theta)}[\log p(x|\theta')]\right] \\
&= -\mathbb{E}_{p(x|\theta)}\left[\nabla_{\theta'}[\log p(x|\theta')]\right] \\
&= -\int p(x|\theta)\nabla_{\theta'}[\log p(x|\theta')]\,dx
\end{aligned}$$

① what's $\nabla_{\theta' \to \theta} \text{KL}[p(x|\theta) \parallel p(x|\theta')]$?

- let $\theta' \to \theta$:

$$
\begin{aligned}
&\nabla_{\theta'} \text{KL}[p(x|\theta) \parallel p(x|\theta')] \; |\theta' \to \theta \\
&= -\int p(x|\theta) \nabla_\theta [\log p(x|\theta)] \, \mathrm{d}x \\
&= -\int p(x|\theta) \frac{\nabla_\theta [p(x|\theta)]}{p(x|\theta)} \, \mathrm{d}x = -\int \nabla_\theta [p(x|\theta)] \mathrm{d}x \\
&= -\nabla_\theta \left[\int p(x|\theta) \mathrm{d}x \right] \\
&= 0
\end{aligned}
$$

② what's $\nabla^2_{\theta' \to \theta} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')]$?

$$
\begin{aligned}
\nabla_{\theta'} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] &= -\int p(x|\theta) \nabla_{\theta'} \log p(x|\theta') \,\mathrm{d}x \\
\implies \nabla^2_{\theta'} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] &= \nabla_{\theta'} \left[-\int p(x|\theta) \nabla_{\theta'} \log p(x|\theta') \,\mathrm{d}x \right] \\
\implies \nabla^2_{\theta' \to \theta} \mathrm{KL}[p(x|\theta) \parallel p(x|\theta')] &= \nabla_{\theta'} \left. \left[-\int p(x|\theta) \nabla_{\theta'} \log p(x|\theta') \,\mathrm{d}x \right] \right|_{\theta' = \theta} \\
&= -\int p(x|\theta) \, \nabla_\theta \left[\nabla_\theta \left[\log p(x|\theta) \right] \right] \mathrm{d}x
\end{aligned}
$$

$$
\begin{aligned}
&\nabla^2_{\theta' \to \theta} \text{KL}[p(x|\theta) \parallel p(x|\theta')] \\
&= -\int p(x|\theta)\, \nabla_\theta \left[\nabla_\theta \left[\log p(x|\theta) \right] \right] \mathrm{d}x = -\int p(x|\theta)\, \nabla_\theta \left[\frac{\nabla_\theta [p(x|\theta)]}{p(x|\theta)} \right] \mathrm{d}x \\
&= -\int p(x|\theta)\, \nabla_\theta \Big[\underbrace{\nabla_\theta [p(x|\theta)]}_{u} \, \underbrace{p(x|\theta)^{-1}}_{v} \Big] \mathrm{d}x \\
&= -\int p(x|\theta) \Big[\underbrace{-\nabla_\theta [p(x|\theta)] p(x|\theta)^{-2} \nabla_\theta [p(x|\theta)]}_{uv'} + \underbrace{\nabla^2_\theta [p(x|\theta)] p(x|\theta)^{-1}}_{u'v} \Big] \mathrm{d}x \quad \text{scalar form} \\
&= -\int p(x|\theta) \left[\nabla^2_\theta [p(x|\theta)] p(x|\theta)^{-1} - \nabla_\theta [p(x|\theta)]^2 p(x|\theta)^{-2} \right] \mathrm{d}x \\
&= -\int p(x|\theta) \left[\frac{\nabla^2_\theta [p(x|\theta)]}{p(x|\theta)} \right] \mathrm{d}x + \int p(x|\theta) \left[\left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right) \left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right)^\top \right] \mathrm{d}x \quad \text{vector-matrix form} \\
&= -\int \nabla^2_\theta [p(x|\theta)] \mathrm{d}x + \mathbb{E}_{p(x|\theta)} \left[\left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right) \left(\frac{\nabla p(x|\theta)}{p(x|\theta)} \right)^\top \right] \\
&= -\nabla^2_\theta \left[\int p(x|\theta) \mathrm{d}x \right] + \mathbb{E}_{p(x|\theta)} \left[\nabla \log p(x|\theta)\, \nabla \log p(x|\theta)^\top \right] \\
&= 0 + \text{F} \\
&= \text{F}
\end{aligned}
$$

Something about Fisher Information (1)

- now, let's have a look at the second derivative:

$$
\begin{aligned}
\nabla^2_{\theta_i,\theta_j}[\log p_\theta(x)] &= \nabla^2_{\theta_i,\theta_j}\left(\frac{\nabla_{\theta_j} p_\theta(x)}{p_\theta(x)}\right) = \nabla_{\theta_i}\left(\frac{\nabla_{\theta_j} p_\theta(x)}{p_\theta(x)}\right) \\
&= \nabla_{\theta_i}\Big(\underbrace{\nabla_{\theta_j} p_\theta(x)}_{u}\,\underbrace{p_\theta(x)^{-1}}_{v}\Big) \\
&= \underbrace{\frac{\nabla^2_{\theta_i,\theta_j} p_\theta(x)}{p_\theta(x)}}_{u'v} \underbrace{- \frac{\nabla_{\theta_i} p_\theta(x)}{p_\theta(x)}\frac{\nabla_{\theta_j} p_\theta(x)}{p_\theta(x)}}_{uv'} \\
\implies \mathbb{E}_{p(x|\theta)}\left[\nabla^2_{\theta_i,\theta_j}[\log p_\theta(x)]\right] &= \mathbb{E}_{p(x|\theta)}\left[\frac{\nabla^2_{\theta_i,\theta_j} p_\theta(x)}{p_\theta(x)}\right] - \mathbb{E}_{p(x|\theta)}\left[\frac{\nabla_{\theta_i} p_\theta(x)}{p_\theta(x)}\frac{\nabla_{\theta_j} p_\theta(x)}{p_\theta(x)}\right] \\
&= 0 - \mathbb{E}_{p(x|\theta)}\left[\nabla_{\theta_i}[\log(p_\theta(x))]\nabla_{\theta_j}[\log(p_\theta(x))]\right] \\
&= 0 - \mathrm{F}_{i,j}
\end{aligned}
$$

Something about Fisher Information (2)

- as a consequence, one may compute:

$$F_{i,j} = \mathbb{E}_{p(x|\theta)}\left[\nabla_{\theta_i}[\log(p_\theta(x))]\nabla_{\theta_j}[\log(p_\theta(x))]\right]$$

or,

$$F_{i,j} = -\mathbb{E}_{p(x|\theta)}\left[\nabla^2_{\theta_i,\theta_j}[\log p_\theta(x)]\right]$$

- of course, we pick the easier of the two!
- now we just proved that,

$$F = \left(\nabla^2_{\theta}\mathrm{KL}[p_{\theta_0} \parallel p_{\theta}]\Big|_{\theta=\theta_0}\right)$$

Final equation: $\text{KL}(\pi \| \beta) = C$

repeat the steps until convergence:

1. feed-forward
2. compute $\nabla_\theta J(\theta_n)$
3. Compute: $\mathrm{F} = \mathbb{E}_{p(x|\theta_n)}\left[\nabla_\theta [J(\theta_n)] \nabla_\theta [J(\theta_n)]^\top\right]$
4. $\theta_{n+1} = \theta_n - \alpha\, \mathrm{F}^{-1} \nabla_{\theta_n} J(\theta_n)$

Then, for policy gradient, we just need to have:

$$
\begin{aligned}
\theta_{n+1} &= \theta_n - \alpha\, \mathrm{F}^{-1} \nabla_\theta \Big(\sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \pi_{\theta_n}(a|s) Q^\pi(s, a) \Big) \\
&= \theta_n - \alpha\, \mathrm{F}^{-1} \Big(\sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \nabla_\theta \log \pi_{\theta_n}(a|s) Q^\pi(s, a) \Big)
\end{aligned}
$$

- $J(\pi_\theta) \equiv J(\theta) = \sum_{s \in \mathcal{S}} d^\pi(s) \sum_{a \in \mathcal{A}} \pi_\theta(a|s) Q^\pi(s, a)$
- if $\tilde{w} = \mathrm{F}^{-1}\nabla_\theta J(\theta)$ is a single natural policy gradient step, then:
- If we can prove $\tilde{w}$ also minimize sqaured error:

$$\tilde{w} = \arg\min_{w} \left(\sum_{s} d^\pi(s) \sum_{a} \pi_\theta(a|s) \left(w^\top \nabla_\theta \log \pi(a|s, \theta) - Q^\pi(s, a) \right)^2 \right)$$

- **interpretation**: Good actions, i.e., those with large $Q^\pi(s, a)$ value should have feature vectors $\nabla_\theta \log \pi(a|s, \theta)$ that have a large inner product with the natural gradient $\tilde{w}$.

Compatible Function Approximation (2)

- We start the reverse: let $\tilde{w}$ minimize sqaured error:

$$\tilde{w} = \arg\min_{w} \left(\sum_s d^\pi(s) \sum_a \pi_\theta(a|s) \big(w^\top \nabla_\theta \log \pi(a|s,\theta) - Q^\pi(s,a) \big)^2 \right)$$

- then,

$$\nabla_w \epsilon(\tilde{w}) = 0$$

$$\nabla_w \epsilon(w) = \nabla_w \left(\sum_s d^\pi(s) \sum_a \pi_\theta(a|s) \big(\nabla_\theta \log \pi_\theta(a|s)^\top w - Q^\pi(s,a) \big)^2 \right)$$

$$\Longrightarrow \sum_s d^\pi(s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) \big(\nabla_\theta \log \pi_\theta(a|s)^\top \tilde{w} - Q^\pi(s,a) \big) = 0$$

$$\Longrightarrow \underbrace{\sum_s d^\pi(s) \sum_a \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top}_{F(\theta)} \tilde{w}$$

$$= \sum_s d^\pi(s) \sum_a \underline{\pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)} Q^\pi(s,a)$$

$$= \underbrace{\sum_s d^\pi(s) \sum_a \underline{\nabla_\theta \pi_\theta(a|s)} Q^\pi(s,a)}_{\nabla_\theta J(\theta)}$$

$$\Longrightarrow F\tilde{w} = \nabla_\theta J(\theta)$$

$$\Longrightarrow \tilde{w} = F^{-1} \nabla_\theta J(\theta)$$

Solve TRPO KL($\pi\|\beta$) $\leq \delta$

- elements of the objective equation:

$$\begin{aligned}
\mathcal{L}_{\theta_k}(\theta) &\approx \underbrace{\mathcal{L}_{\theta_k}(\theta_k)}_{0} + g^\top(\theta - \theta_k) \\
&= g^\top(\theta - \theta_k) \qquad \text{where } g = \nabla_\theta \mathcal{L}_{\theta_k}(\theta) \mid \theta_k \\
\bar{\text{KL}}(\theta\|\theta_k) &\approx \underbrace{\bar{\text{KL}}(\theta_k\|\theta_k)}_{0} + \underbrace{\nabla_\theta \bar{\text{KL}}(\theta_k\|\theta_k)}_{0} + \frac{1}{2}(\theta - \theta_k)^\top \text{F}(\theta - \theta_k) \\
&= \frac{1}{2}(\theta - \theta_k)^\top \text{F}(\theta - \theta_k) \qquad \text{where F} = {\nabla_\theta}^2 \bar{\text{KL}}(\theta\|\theta_k) \mid \theta_k
\end{aligned}$$

Objective function

- objective function of:

$$\max_{\pi} \underbrace{\left[\mathbb{E}_{\tau \sim \beta} \left[\sum_{t=0}^{\infty} \gamma^t \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t) \right] \right]}_{\mathcal{L}_{\theta_k}(\theta)}$$

$$\text{s.t.} \quad \text{KL}(\pi \| \beta) \leq \delta$$

can be re-formulated as:

$$\theta_{k+1} = \arg\max_{\theta} \left[g^{\top}(\theta - \theta_k) \right]$$

$$\text{s.t.} = \frac{1}{2}(\theta - \theta_k)^{\top} \text{F}(\theta - \theta_k) \leq \delta$$

- answer:

$$\theta_{k+1} = \theta_k + \frac{1}{\sqrt{g^{\top} \text{F}^{-1} g}} \text{F}^{-1} g$$

KKT condition

Let $x \equiv (\theta - \theta_k)$:

- primal:

$$f = \max \left[g^\top x \mid \frac{1}{2} x^\top \mathrm{F} x \leq \delta, \quad x, c \in \mathbb{R}^n, \quad \mathrm{F} \in \mathbb{R}^{n \times n} \right]$$

- Lagrangian

$$\begin{aligned} \mathcal{L}(x, \lambda) &= -g^\top x + \lambda \frac{1}{2} (x^\top \mathrm{F} x - 2\delta) \\ \implies \nabla_x \mathcal{L}(x, \lambda) &= -g + \lambda \mathrm{F} x \end{aligned}$$

- **KKT conditions**:

$$-g + \lambda \mathrm{F} x = 0, \quad \lambda \geq 0, \quad \lambda (x^\top \mathrm{F} x - 2\delta) = 0, \quad x^\top \mathrm{F} x \leq 2\delta$$

find x^*

- condition $\lambda(x^\top Fx - 2\delta) = 0$ states two cases: if $x^\top Fx < 2\delta \implies \lambda = 0$, and from condition $-g + \lambda Fx = 0 \implies g = 0$, which can **not** be the max
  Hence we take another case: $\lambda > 0$, $x^\top Hx = 2\delta$
- find expression of λ without having x

$$
\begin{aligned}
-g + \lambda Fx = 0 \quad &\implies x = \frac{1}{\lambda} F^{-1} g \\
x^\top Fx &= \left(\frac{1}{\lambda} F^{-1} g\right)^\top F \left(\frac{1}{\lambda} F^{-1} g\right) \\
&= \frac{1}{\lambda^2} g^\top \underbrace{F^{-1}}_{\text{symmetric}} FF^{-1} g = \frac{1}{\lambda^2} g^\top F^{-1} g = 2\delta \\
&\implies \lambda^2 = \frac{g^\top F^{-1} g}{2\delta} \\
&\implies \lambda = \sqrt{\frac{g^T F^{-1} g}{2\delta}} \qquad \text{since} \lambda \geq 0
\end{aligned}
$$

- substitute λ in the expression of x:

$$
x^* = \frac{1}{\lambda} F^{-1} g = \sqrt{\frac{2\delta}{g^T F^{-1} g}} F^{-1} g
$$

Gradient descend via finding maximum first

- solving it using:

$$
\begin{aligned}
x \equiv (\theta - \theta_k) &\implies x^* \equiv (\theta_{k+1} - \theta_k) \\
\implies \theta_{k+1} = \theta_k + &\sqrt{\frac{2\delta}{\hat{g}_k \hat{\mathrm{F}}_k^{-1} \hat{g}_k}} \hat{\mathrm{F}}_k^{-1} \hat{g}_k
\end{aligned}
$$

- $\hat{\mathrm{F}}_k^{-1}$ is too computational! but we don't need to compute it, however, we can compute $\hat{\mathrm{F}}_k^{-1} \hat{g}_k$ together!

Conjugate Gradient Descend - why need conjugate?

- we have a 2-d function $f(x_1, x_2)$:
- suppose step k occurred along x_1-axis, and led to position $\mathbf{x}^{k+1}$
- at $\mathbf{x}^{k+1}$, $f(\mathbf{x}^{k+1})$ is minimized in its x_1 component:

$$\frac{\partial f(\mathbf{x}^{k+1})}{\partial x_1} = 0$$

- next step is along x_2-axis: that step leads to a position $\mathbf{x}^{k+2}$: we find the approprate step, such that:

$$\frac{\partial f(\mathbf{x}^{k+2})}{\partial x_2} = 0$$

- we know $\frac{\partial^2 f(\mathbf{x}^{k+2})}{\partial x_1 \, \partial x_2} = \frac{\partial}{\partial x_2}\left(\frac{\partial f(\mathbf{x}^{k+2})}{\partial x_1}\right)$, then:

$$\frac{\partial^2 f(\mathbf{x}^{k+2})}{\partial x_2 \partial x_1} \neq 0 \implies \frac{\partial f(\mathbf{x}^{k+2})}{\partial x_1} \neq 0$$

- in words, it says if $\mathbf{x}^{k+2}$ is **not** overall stationery/saddle point, and we also know $\mathbf{x}^{k+2}$ **is** stationery point in x_2 direction; then it **mustn't** be stationery point in x_1 direction
- we want to move along direction other than x_2-axis, such that $\frac{\partial f(\mathbf{x}^{k+2})}{\partial x_1}$ remains zero

Q-conjugate

- we need to search for new non-axis directions:
- $\{d_1, d_2, \dots, d_n\}$ are said to be Q-conjugate, such that,

$$d_j^\top Q d_k = 0 \qquad j \neq k$$

- when Q is also symmetric, $\{\lambda_k, v_k\}$ are eigen-(value,vector) pair, we know all eigen-vectors are orthogonal:

$$\begin{aligned} Qv_k &= \lambda_k v_k \\ \implies \quad v_j^\top Q v_k &= \lambda_k v_j^\top v_k = 0 \qquad j \neq k \end{aligned}$$

- so eigen-vectors $\{v_1, \dots v_n\}$ of symmetric matrix can be thought as special case of Q-conjugate vectors, where these vectors are ortho-normal without Q

CGD: Linear independence

- let Q be **positive definite**, then all its Q-conjugate vectors $\{d_1, d_2, \dots, d_n\}$ are linearly independent
- **proof by contradiction**, i.e., suppose one of its vector say d_k can be written in linear combination of $d_1, \dots, d_{k-1}$:

$$\begin{aligned} d_k &= \alpha_1 d_1 + \cdots + \alpha_{k-1} d_{k-1} \\ \implies d_k^\top Q d_k &= d_k^\top Q \left(\alpha_1 d_1 + \cdots + \alpha_{k-1} d_{k-1}\right) \\ &= d_k^\top Q \alpha_1 d_1 + \cdots + d_k^\top Q \alpha_{k-1} d_{k-1} \\ &= 0 \end{aligned}$$

contradiction part is, by definition of positive definiteness: $d_k^\top Q d_k > 0 \; \forall d_k \neq 0$!

compute α_k independently

- if we are to minimize a **quadratic** problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} - b^\top \mathbf{x} + c$$

- if matrix $Q \in \mathbb{R}^{n\times n}$ is positive definite, then minimal value $\mathbf{x}^*$ is:

$$Q\mathbf{x}^* = b$$

- let $\{d_0, d_1, \dots, d_{n-1}\}$ be arbitary Q-conjugate set

$$\begin{aligned}
\mathbf{x}^* &= \alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1} && \text{linearly-independent basis} \\
\Longrightarrow \quad d_k^\top Q\mathbf{x}^* &= d_k^\top Q\left(\alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1}\right) && \times \text{ by arbitrary } k^{\text{th}} \\
&= \alpha_k d_k^\top Q d_k \\
\Longrightarrow \quad \alpha_k &= \frac{d_k^\top Qx^*}{d_k^\top Q d_k} = \frac{d_k^\top b}{d_k^\top Q d_k}
\end{aligned}$$

- **beauty** is that we don't need to know $\mathbf{x}^*$ to compute α_k, only Q-conjugacy is required

Conjugate Direction

$$
\begin{aligned}
\mathbf{x}^* &= \alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1} \\
&= \sum_{k=0}^{d-1} \frac{d_k^\top b}{d_k^\top Q d_k} d_k \qquad\qquad \text{substitute } \alpha_k = \frac{d_k^\top b}{d_k^\top Q d_k}
\end{aligned}
$$

- the above can be achieved in parallel, where each d_k does **not** minimizing anything
- also it is **not** an algorithm, it simply decomposes $\mathbf{x}^*$
- instead, we try to solve along a **path**, with an initial point $\mathbf{x}^0$:

$$
\begin{aligned}
\mathbf{x}_1 &= \color{red}{\mathbf{x}_0} + \alpha_0 d_0 \\
&\cdots \\
\mathbf{x}_k &= \color{red}{\mathbf{x}_0} + \alpha_0 d_0 + \cdots + \alpha_{k-1} d_{k-1} \\
&\cdots \\
\mathbf{x}^* &= \color{red}{\mathbf{x}_0} + \alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1}
\end{aligned}
$$

- what about the new α_k to match with this path?

now we have $\mathbf{x}_0$

- $\mathbf{x}_0 \in \mathbb{R}^n$ be an arbitrary starting point:
- so instead of writing $\mathbf{x}^* = \sum_{k=0}^{d-1} \alpha_k d_k$
- we also know $g_k \equiv \nabla f(\mathbf{x}_k) = Q\mathbf{x}_k - b = Q\mathbf{x}_k - Q\mathbf{x}^* = Q(\mathbf{x}_k - \mathbf{x}^*)$
- instead of decompose $\mathbf{x}^*$, let's now try to decompose $\mathbf{x}^* - \mathbf{x}_0$:

$$\mathbf{x}_1 - \mathbf{x}_0 = \underbrace{\mathbf{x}_0 + \alpha_0 d_0}_{\mathbf{x}_1} - \mathbf{x}_0$$

$$\mathbf{x}_k - \mathbf{x}_0 = \underbrace{\mathbf{x}_0 + \alpha_0 d_0 + \cdots + \alpha_k d_{k-1}}_{\mathbf{x}_k} - \mathbf{x}_0 = \alpha_0 d_0 + \cdots + \alpha_{k-1} d_{k-1}$$

$$\mathbf{x}^* - \mathbf{x}_0 = \underbrace{\mathbf{x}_0 + \alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1}}_{\mathbf{x}^*} - \mathbf{x}_0 = \alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1}$$

$$\begin{aligned}
\implies d_k^\top Q(\mathbf{x}^* - \mathbf{x}_0) &= d_k^\top Q\left(\alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1}\right) \\
&= d_k^\top Q \alpha_k d_k \\
\implies \alpha_k &= \frac{\color{red}{d_k^\top Q(\mathbf{x}^* - \mathbf{x}_0)}}{d_k^\top Q d_k} \\
&= -\frac{d_k^\top g_0}{d_k^\top Q d_k}
\end{aligned}$$

- **recap**, for $\mathbf{x}^* = \alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1}$:

$$\mathbf{x}^* = \sum_{k=0}^{d-1} \underbrace{\frac{d_k^\top b}{d_k^\top Q d_k}}_{\alpha_k} d_k$$

- **recap**, for $\mathbf{x}^* = \alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1} + \mathbf{x}_0$:

$$\mathbf{x}^* - \mathbf{x}_0 = \sum_{k=0}^{d-1} \underbrace{-\frac{d_k^\top Q(\mathbf{x}^* - \mathbf{x}_0)}{d_k^\top Q d_k}}_{\alpha_k} d_k$$

$$\mathbf{x}^* = \sum_{k=0}^{d-1} \underbrace{-\frac{d_k^\top Q(\mathbf{x}^* - \mathbf{x}_0)}{d_k^\top Q d_k}}_{\alpha_k} d_k + \mathbf{x}_0$$

- we will see that to write α_k in terms of $Q(\mathbf{x}^* - \mathbf{x}_0)$ may **not** be as useful as to write in terms of $\mathbf{x}_k$

Expanding subspace theorem

- looking at:

$$\begin{aligned} d_k^\top Q(\mathbf{x}^* - \mathbf{x}_0) &= d_k^\top Q(\mathbf{x}^* - \mathbf{x}_k + \mathbf{x}_k - \mathbf{x}_0) = d_k^\top Q(\mathbf{x}^* - \mathbf{x}_k) + d_k^\top Q(\mathbf{x}_k - Q\mathbf{x}_0) \\ &= d_k^\top Q(\mathbf{x}^* - \mathbf{x}_k) + d_k^\top Q\left(\alpha_0 d_0 + \cdots + \alpha_{n-1} d_{n-1}\right) \\ &= d_k^\top Q(\mathbf{x}^* - \mathbf{x}_k) \end{aligned}$$

- noted that $d_k^\top Q(\mathbf{x}^* - \mathbf{x}_0) = d_k^\top Q(\mathbf{x}^* - \mathbf{x}_k) \not\Rightarrow Q(\mathbf{x}^* - \mathbf{x}_0) = Q(\mathbf{x}^* - \mathbf{x}_k)$
- think about the case:

$$[1 \quad 1]\, v_1 = [1 \quad 1]\, v_2 = 5 \quad \text{but} \quad v_1 = [4 \quad 1] \text{ and } v_2 = [1 \quad 4] \text{ satisfy}$$

- therefore:

$$\alpha_k = \frac{d_k^\top Q(\mathbf{x}^* - \mathbf{x}_0)}{d_k^\top Q d_k} = -\frac{d_k^\top g_0}{d_k^\top Q d_k} = \frac{d_k^\top Q(\mathbf{x}^* - \mathbf{x}_k)}{d_k^\top Q d_k} = -\frac{d_k^\top g_k}{d_k^\top Q d_k}$$

- **recap**: we move from $\mathbf{x}_0$ by adding Q-conjugate directions $\{d_1, \ldots d_n\}$, each time by $\alpha_k = -\frac{d_k^\top g_k}{d_k^\top Q d_k}$ amount
- we need to prove why this movement is getting “better”, i.e., each k step **minimizes all previous directions**

Looking at the algorithm closely

- to know if $\mathbf{x}_k$ is minimizing dimensions along its path using step size $\alpha_k = -\frac{d_k^\top g_k}{d_k^\top Q d_k}$:

$$\mathbf{x}_k \xrightarrow{\alpha_k \times d_k} \mathbf{x}_{k+1} \qquad\qquad \mathbf{x}_{k+1} \xrightarrow{\alpha_{k+1} \times d_{k+1}} \mathbf{x}_{k+2}$$

  where each $\mathbf{x}_k$ is used to compute its corresponding $g_k \equiv \nabla(\mathbf{x}_k)$
- starting in the first step, given arbitrary point $\mathbf{x}_0$:

$$\begin{aligned}\mathbf{x}_1 &= \mathbf{x}_0 + \alpha_0 d_0 \\ g_0 &= Q\mathbf{x}_0 - b\end{aligned}$$

- obviously, we hope $\mathbf{x}_1$ to minimize the **line** (direction) $\mathbf{x}_0 + \alpha_0 d_0$
- this is equivalently saying, $g_1 \equiv \nabla f(\mathbf{x}_1) \perp (\mathbf{x}_0 + \alpha_0 d_0)$
- think this way, we now have changed the coordinates from one ortho-normal basis to another: $[x_1, x_2] \rightarrow [u, v]$ let:

$$(u = (\mathbf{x}_0 + \alpha_0 d_0) \quad \text{and} \quad v \perp u) \implies \left[\frac{\partial f}{\partial u}, \frac{\partial f}{\partial v}\right] = \left[0, \frac{\partial f}{\partial v}\right]$$

Looking at the algorithm closely

- we have,

$$
\begin{aligned}
g_1 &= \nabla f(\mathbf{x}_1) = Q\mathbf{x}_1 - b \\
&= Q(\mathbf{x}_0 + \alpha_0 d_0) - b = (Q\mathbf{x}_0 - b) + \alpha_0 Q d_0 \\
&= g_0 + \alpha_0 Q d_0
\end{aligned}
$$

- $g_1 \not\perp d_0$ in general, but we can show a particular choice α_0 makes it do, i.e., x_1 minimizes the line $\mathbf{x}_0 + \alpha_0 d_0$

$$
\begin{aligned}
d_0^\top g_1 &= d_0^\top g_0 + d_0^\top \alpha_0 Q d_0 && \times d_0^\top \text{ on each side} \\
&= d_0^\top g_0 + \alpha_0 d_0^\top Q d_0 \\
&= d_0^\top g_0 - \frac{d_0^\top g_0}{d_0^\top Q d_0} d_0^\top Q d_0 && \text{sub } \alpha_0 = -\frac{d_0^\top g_0}{d_0^\top Q d_0} \\
&= d_0^\top g_0 - d_0^\top g_0 = 0 \\
\implies & d_0 \perp g_1
\end{aligned}
$$

- above shows the choice d_0 is also somewhat arbitrary
- **to understand** by choose a different $\mathbf{x}_0$, results a different g_0, having an arbitrary (g_0, d_0) pair results a unique $\alpha_0 = -\frac{d_0^\top g_0}{d_0^\top Q d_0}$ making $\mathbf{x}_1$ the minimum of the **line** $\mathbf{x}_0 + \alpha_0 d_0$
- however, a sensible choice is $d_0 = -\nabla f(\mathbf{x}_0) = -g_0$

Expanding Subspace Theorem

- knowing $g_1 \perp d_0$, we also can prove similarly that:

$$g_k \perp \text{span}(\underbrace{d_0, \dots, d_{k-1}}_{k \text{ terms}})$$

for example, if $\mathbf{x}_2 \perp (\mathbf{x}_0 + \alpha_0 d_0)$ and $\mathbf{x}_2 \perp (\mathbf{x}_1 + \alpha_1 d_1)$, we know that $\mathbf{x}_2 \perp$ a **surface** span of the two perpendicular lines d_0 and d_1, we write this as:

$$g_2 \perp \text{span}(\underbrace{d_0, d_1}_{2 \text{ terms}})$$

we can drop $\mathbf{x}_0$ and $\mathbf{x}_1$

- we can see that $\mathbf{x}_k$ minimizes f over $\{\mathbf{x}_0 + \text{span}(d_0, \dots, d_{k-1})\}$
- therefore, it's obvious"

$$\mathbf{x}_n = \underset{\mathbf{x} \in \{\mathbf{x}_0 + \text{span}(d_0, \dots, d_{n-1})\}}{\arg\min} \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} - b^\top \mathbf{x}$$

Determine directions

- one more thing missing, we know it works well for any arbitrary Q-conjugate vectors $\{d_0, \dots, d_n\}$:
- a sensible guess of d_1 would be (we already used $d_0 = -\nabla f(\mathbf{x}_0) = -g_0$:

$$d_1 = -\nabla f(\mathbf{x}_1) + \beta_0 d_0 = -g_1 + \beta_0 d_0$$

- use definition of conjugacy:

$$\begin{aligned} d_1^\top Q d_0 &= 0 \\ \implies \; (-g_1 + \beta_0 d_0)^\top d_0 &= 0 \\ -g_1^\top Q d_0 + \beta_0 d_0^\top Q d_0 &= 0 \\ \beta_0 &= \frac{g_1^\top Q d_0}{d_0{}^\top Q d_0} \end{aligned}$$

Conjugate Gradient Algorithm

1. let f be a quadratic function:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top Q\mathbf{x} + b^\top \mathbf{x} + c$$

2. **initialize**: Let $i = 0$ and $\mathbf{x}_i = \mathbf{x}_0$, $d_i = d_0 = \nabla f(\mathbf{x}_0)$
3. compute α_0 to minimize the function $f(\mathbf{x}_i + \alpha d_i)$:

$$\begin{aligned}\alpha_k &= -\frac{{d_k}^\top (Q\mathbf{x}_k + b)}{{d_k}^\top Q d_k} \\ &= -\frac{d_k^\top g_k}{d_k^\top Q d_k}\end{aligned}$$

4. update

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha d_k \qquad g_k = Qx_k - b$$

5. update the direction:

$$d_{k+1} = -g_{k+1} + \beta_k d_k \qquad \beta_k = \frac{g_{k+1}^\top Q d_k}{{d_k}^\top Q d_k}$$

6. Repeat steps 2-4 until we have looked in n directions, where $\mathbf{x} \in \mathbb{R}^n$

How to find $\bar{\mathrm{F}}\bar{g}$?

- how does it translate to our problem, i.e.,

$$\theta_{k+1} = \theta_k + \sqrt{\frac{2\delta}{\hat{g}_k \hat{\mathrm{F}}_k^{-1} \hat{g}_k}} \hat{\mathrm{F}}_k^{-1} \hat{g}_k$$

- if matrix $Q \in \mathbb{R}^{n \times n}$ is positive definite, then minimal value $\mathbf{x}^*$ is:

$$Q\mathbf{x}^* = b \implies \quad x^* = Q^{-1} b$$

- as per CGA algorithm, which requires computation of Qd_k, or $\bar{\mathrm{F}}\bar{g}_k$ (note, not $\bar{\mathrm{F}}^{-1}\bar{g}$)
- **Direct method** can help with it:

$$\mathrm{F}_{ij} = \frac{\partial}{\partial \theta_j} \frac{\partial \mathrm{KL}}{\partial \theta_i}$$

$$\begin{aligned} f_k &= \sum_j \mathrm{F}_{kj} g_j = \sum_j \frac{\partial}{\partial \theta_j} \frac{\partial \mathrm{KL}}{\partial \theta_k} g_j = \left(\frac{\partial}{\partial \theta} \frac{\partial \mathrm{KL}}{\partial \theta_k} \right)^\top g \\ &= \frac{\partial}{\partial \theta_k} \underbrace{\sum_j \frac{\partial \mathrm{KL}}{\partial \theta_j} g_j}_{\text{scalar}} = \frac{\partial}{\partial \theta_k} \underbrace{\left(\frac{\partial \mathrm{KL}}{\partial \theta} \right)^\top g}_{\text{scalar}} \end{aligned}$$

Proximal Policy Optimization (PPO)

- TRPO is expressed as:

$$\max_{\pi} \left[\mathbb{E}_{\tau\sim\beta} \left[\sum_{t=0}^{\infty} \gamma^t \frac{\pi(a_t|s_t)}{\beta(a_t|s_t)} A^{\beta}(s_t, a_t) \right] - C\sqrt{\mathbb{E}_{s\sim d_k^{\beta}}[\mathrm{KL}(\pi\|\beta)[s]]} \right]$$

- PPO is expressed as, using $r_t(\theta) = \frac{\pi_{\theta}(a|s)}{\beta(a_t|s_t)}$:

$$\max_{\pi} \left[\mathbb{E}_{\tau\sim\beta} \left[\sum_{t=0}^{\infty} \gamma^t \min\left(\underbrace{r_t(\theta)A^{\beta}(s_t, a_t)}, \underbrace{\mathrm{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)A^{\beta}(s_t, a_t)} \right) \right] \right]$$

- if $r_t(\theta)$ falls outside $(1-\epsilon)$ and $(1+\epsilon)$, $A^{\beta}(s_t, a_t)$ will be clipped
- sign of $A^{\beta}(s_t, a_t)$ plays a part:
  1. if $A^{\beta}(s_t, a_t) > 0$, PPO clips at $r_t(\theta) = 1+\epsilon$
  2. if $A^{\beta}(s_t, a_t) < 0$, PPO clips at $r_t(\theta) = 1-\epsilon$
- Therefore PPO is **not** the same as:

$$\max_{\pi} \left[\mathbb{E}_{\tau\sim\beta} \left[\sum_{t=0}^{\infty} \gamma^t \min\left(\underbrace{r_t(\theta)}, \underbrace{\mathrm{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)} \right) A^{\beta}(s_t, a_t) \right] \right]$$